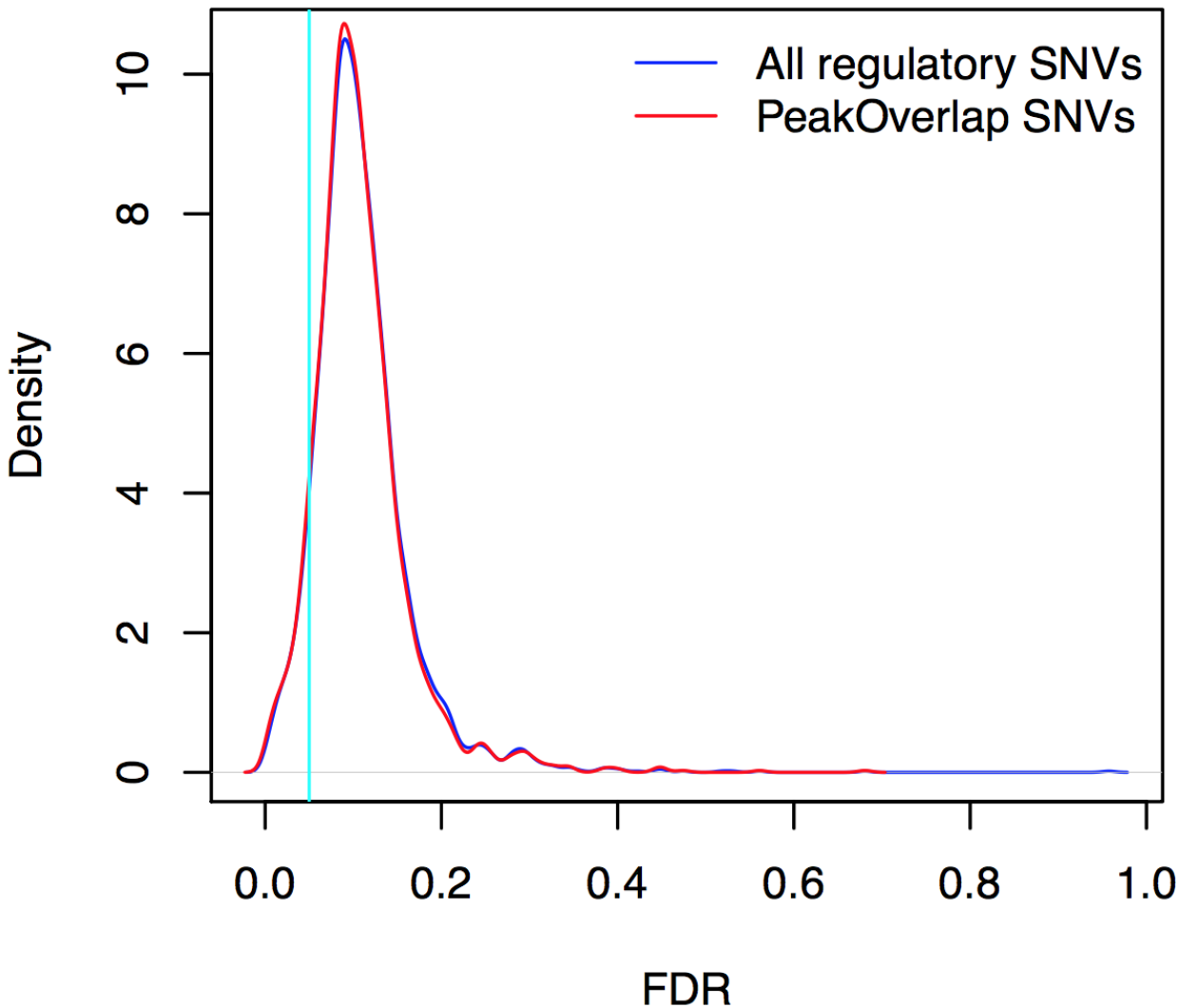
Motivation

Skewed distribution of FDR indicates it difficult to define negative set using a cutoff.
If the data with fdr <=0.05 like positive dataset, the remaing like unlabeled dataset with positive and negative mixed data.
We then define the probablem as a **positive unlabeled learning problem**



Semi-supervised learning

One class SVM (Bernhard et. al. MIT Press (2000) train model using positive data, to separates all the data points from the origin (in feature space F) and maximizes the distance from this hyperplane to the origin.
A binary class is defined to capture regions in the input space where the probability density of the data lives.

Then use learned model to predict unlabeled data set; Positive and negative predicted label data are combined for futher learning

LASSO

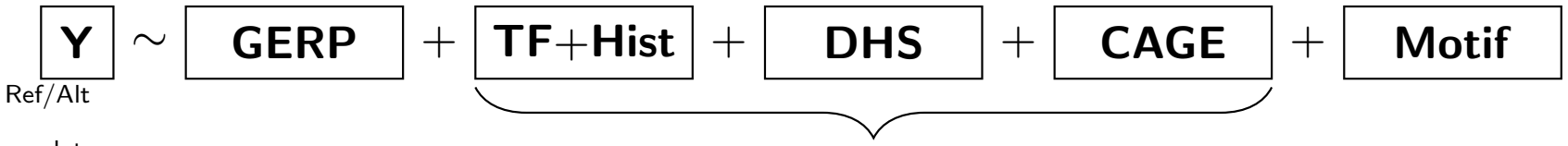
The positive and negative dataset defined using semi-supervised learning, are further trained using LASSO.
The Cost function with L1-norm:
$J(\theta) = -\frac{1}{n}\sum(ylog(h_{(\theta)}(x)) + (1-y)log(1-h_{(\theta)}(x))) + \lambda||\theta||$

where
$h_{(\theta)}(x) = \frac{1}{1+e^{-(\theta^T x+\theta_0)}}$

cross validation help to find best sub feature set with highest AUC (0.82), and AUC within 1st SE but much simple model.

Use the θ (coefficient) learned using 1st SE (avoid overfitting) model to predict each positions for 403 regions

Model and features



Ref/Alt regulatory change

Tissue specific binary feature

Train use Gm12878
TF,Hist, CAGE and DHS

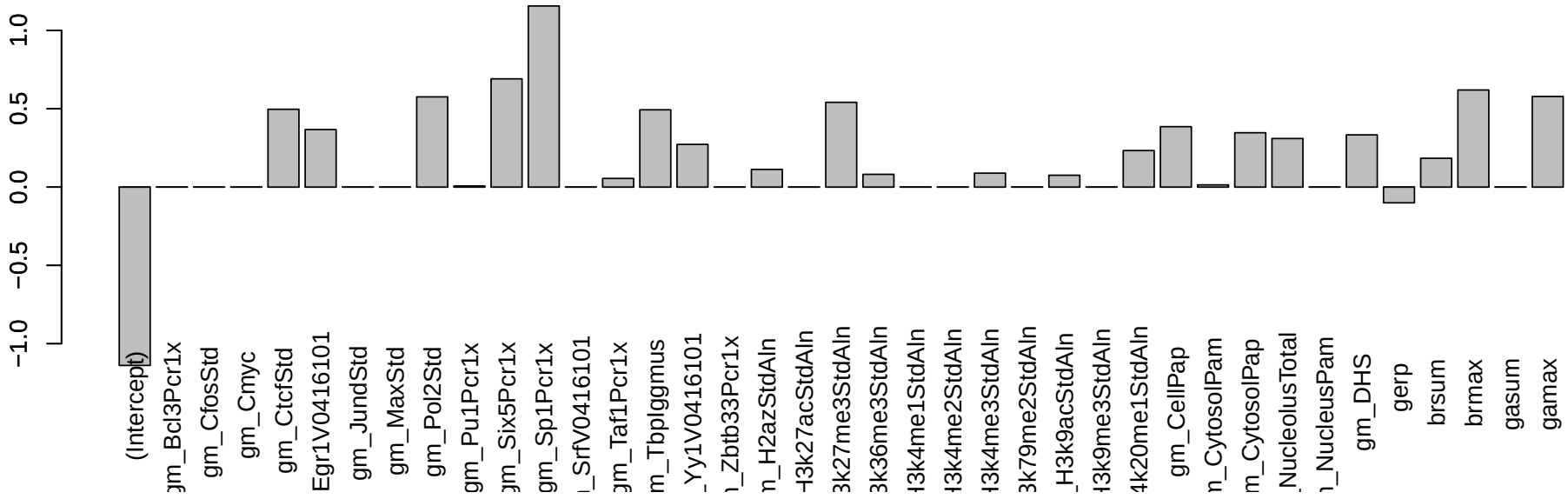
Predict use K562
TF,Hist, CAGE and DHS

34 strong tissue specific SNV identified, 4 gain ref/alt change

Whether it worth to test these 4 SNVs

predict SNV for 403 regions, selected based on Nature Biotch paper. EAch region at least have 2 SNVs with high and low prob

Coef of selected features



$$y\prime = \frac{1}{1+e^{-\tilde{\theta}^T x}}$$